Exploring the Boundaries of Plausibility: Empirical Study of a Key Problem in the Design of Computer-Based Clinical Simulations

Charles P. Friedman, PhD[1], Guido G. Gatti[1], Gwendolyn C. Murphy, PhD[2], Timothy M. Franz, PhD[3], Paul L. Fine, MD[4], Paul S. Heckerling, MD[5], Thomas M. Miller, MD[6]

[1]Center for Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA
[2]Department of Community and Family Medicine, Duke University, Durham, NC
[3]Deparment of Psychology, St. John Fisher College, Rochester, NY
[4] Department of Medicine, University of Michigan, Ann Arbor, MI
[5]Department of Medicine, University of Illinois, Chicago, IL
[6]Department of Medicine, University of North Carolina, Chapel Hill, NC

ABSTRACT

All clinical simulation designers face the problem of identifying the plausible diagnostic and management options to include in their simulation models. This study explores the number of plausible diagnoses that exist for a given case, and how many subjects must work up a case before all plausible diagnoses are identified. Data derive from 144 residents and faculty physicians from 3 medical centers, each of whom worked 9 diagnostically challenging cases selected from a set of 36. Each subject generated up to 6 diagnostic hypotheses for each case, and each hypothesis was rated for plausibility by a clinician panel. Of the 2091 diagnoses generated, 399 (19.1%), an average of 11 per case, were considered plausible by study criteria. The distribution of plausibility ratings was found to be statistically case dependent. Averaged across cases, the final plausible diagnosis was generated by the 28th clinician (sd = 8) who worked the case. The results illustrate the richness and diversity of human cognition and the challenges these pose for creation of realistic simulations in biomedical domains.

INTRODUCTION

A growing literature attests to the importance of simulation technology in the pre-professional and continuing education of clinicians across the health professions [1-3]. Entire conferences are now being devoted to this topic [4]. As documented in this literature, well-designed, high-fidelity simulations provide limitless opportunities for practice of cognitive tasks and procedural skills, allowing learners to make mistakes on machines instead of human patients or animals. Simulations also provide specific feedback to guide learning, and can be used for competence assessment to promote patient safety and reduce medical errors.

Since it is not possible to capture the full extent of reality in any simulator, and especially in clinical medicine, every simulation designer faces a fundamental problem of what to include and what to exclude from the models and user options that power a simulation. Error in the direction of providing too much can make simulations almost impossibly expensive to produce and slow to execute, with no material benefit to the educational or assessment process. Error in the direction of providing too little can make simulations unrealistic, by overly limiting learners' options to act.

More specifically, the simulation model must make available to learners all reasonable diagnostic and management options, explicitly including and "understanding" every action that the learner can take, both correct and incorrect, in the course of working through a case problem. Failing to open all plausible options to the learner can result in cueing that destroys much of the realism of the simulation. For example, a student working on a simulated case that is really a case of pneumonia may want to order studies appropriate to a workup for lung cancer. If the simulation model does not include the option of ordering tests specific to a cancer workup, it is cueing the learner to the fact that the problem is not cancer. This interjects unrealism that can destroy the educational value of a simulation. (This is a different issue from *intentionally* cueing a novice student to the fact that the problem is not cancer, in order to help them as they work through the simulation.)

These dilemmas for simulation designers exist whether the interface of the simulation emphasizes forced choices from picklists or open-ended text entry by the user/learner. In the former instance, the simulation model must include all plausible options in the picklists. Too few options engenders cueing; too many may be unacceptably cumbersome. In the latter instance of open-ended text entry, the simulation must parse and recognize all reasonable entries by the learner and allow the learner's requested action to be taken. These dilemmas are not new; they reach back to the earliest work in this area [5-7].

So how, then, does a simulation designer determine what is plausible? How does the designer include enough options to make the simulation realistic without going to the time and expense of including extraneous material that will make the simulation overly cumbersome and ornate? One approach is a prospective empirical study, as part of the simulation design process, where subjects are stepped through the case and, at various points, asked to "think out loud" regarding their current diagnostic assessment or the actions they might take at that point. This approach would, over a series of subjects, generate a range of hypotheses and actions, eventually identifying those that are appropriate to include in the finished simulation design. But how many subjects are required in such a study before one can be confident that everything plausible has arisen? Can simulation designers rely on their own judgments of what is plausible to include as options for the learners?

This study offers some initial empirical evidence on the bounds of plausibility in the domain of diagnosis in internal medicine, which was the domain of some of the earliest work on simulation in medical education [7]. We explore the following specific research questions:

1. For challenging diagnostic problems in internal medicine, how many plausible diagnoses exist? What is the mean number of plausible diagnoses per case and do these distributions vary across cases?

2. Anticipating formal studies to identify all the plausible options for simulated cases, how many subjects are needed to identify the full extent of plausibility? How does this result change if simulation designers are satisfied to identify less than the full set of plausible hypotheses?

METHODS

To address these questions, we employed a large dataset originally collected for a study of the impact of decision support systems on the accuracy of clinicians' diagnoses [8]. We developed for the original study detailed written synopses of 36 diagnostically challenging cases from patient records at the University of Illinois at Chicago, the University of Michigan, and the University of North Carolina. Each institution contributed twelve cases with firmly established final diagnoses. The 2-4 page case synopses were designed to provide a complete portrayal of the patient. As such, they contained comprehensive historical, examination and diagnostic test information. The case descriptions did not, however, contain results of "definitive" or pathognomic tests that would have made the correct diagnosis obvious to most or all clinicians. We divided the 36 cases into four sets of 9 cases each. The sets were balanced for pathophysiologic process, degree of difficulty, and institution of origin.

We then recruited to the study 216 subjects from these same institutions: 72 fourth year medical students, 72 second- and third-year internal medicine residents, and 72 general internists with faculty appointments and at least two years of post-residency experience. (They averaged 11 years of experience.) Recruitment was balanced so that each institution contributed 24 subjects at each level of experience. Each subject was randomly assigned to work the 9 cases comprising one of the 4 case sets, so each case was completed by 18 students, 18 residents and 18 faculty physicians. Each subject worked through each of the assigned cases first without, and then with, assistance from an assigned computer-based decision support system. On each pass through each case, subjects generated a differential diagnosis consisting of up to 6 ordered diagnostic hypotheses.

The data addressing the research questions in this study emanated from the first pass through each case by the subjects, so the results in this study are a reflection of human cognition unaided by computer-based or other decision support. Also, because we discovered that the medical students were largely overmatched by these cases--they diagnosed only 26% of cases correctly and often were guessing rather than providing knowledge-directed hypotheses--we based the analyses reported here on the performance of the resident and faculty subjects. Eliminating the medical students from this analysis reduced the total number of subjects in the study to 144, and reduced from 54 to 36 the number of subjects who worked each case.

As part of the generation of the complete dataset for the study, a panel of three experienced internists (co-authors PLF, PSH, TMM) rated the plausibility of each diagnosis reported in the hypothesis lists. Plausibility was judged in the context of each case. Plausibility of each hypothesis for each case, was rated by each judge on a 1-7 scale, with a score of "7" given to the correct hypothesis, for the case, and "1" to a completely improbable hypothesis. We averaged the ratings of the panelists to compute a plausibility score for each unique diagnostic hypothesis within each case. The averaging generated non-integer plausibility scores for some hypotheses. Specific diagnostic hypotheses that recurred across multiple cases typically received different plausibility scores in the context of these different cases. Interjudge reliability of the plausibility ratings of these hypotheses averaged .85 across the cases [9].

Analyses addressing both research questions required us to explore thresholds for considering diagnoses to be "plausible". Based on the semantics of the 7-point scale used for the ratings, consensus of the research team was that any hypothesis with a plausibility score less than 4 should not be considered "plausible", and that any hypothesis with a score of 5 or higher should be considered "plausible". We elected, for this initial exploration, to employ an intermediate threshold of plausibility, considering as plausible within each case all diagnostic hypotheses with a score greater than 4.

To address the first research question regarding the number and distribution of plausible hypotheses per case, we created a 36 x 7 contingency table with a row for each case and a column for each plausibility level, rounding each score to the corresponding integer value. The table cells contained the counts of hypotheses at each plausibility level for each case. We computed the mean and distributional features, for each case and across cases, of the numbers of plausible diagnoses (those with plausibility score greater than 4). We then tested statistically the case-dependence of these distributions using chi-square analysis of the contingency table.

Addressing the second research question, how many subjects are required to identify the full set of plausible hypotheses, necessitated that we make our findings statistically independent of any particular ordering of subjects in the dataset. To this end, we generated for each case 100 random orderings of subjects. For each ordering we computed the serial number of the subject who generated the last of the set of plausible hypotheses for that case; i.e. the subject for that ordering after whom no more plausible hypotheses were seen. The average of these serial numbers across the orderings created an unbiased estimate of the mean number of subjects required to generate all plausible hypotheses, as identified by the full group of subjects, for that case. If, for example, the plausible hypotheses tended to recur in most subjects' lists, we would expect that many fewer than the full group of subjects would be required to identify the full set. We then repeated the analysis to determine the serial numbers of the subjects who generated the next-to-last plausible diagnosis, for each case, as well as the serial numbers of the subjects who generated the last plausible diagnosis but two.

RESULTS

Research Question 1: The faculty and resident subjects collectively generated 2091 diagnostic hypotheses over all cases, or 58 unique diagnoses on average per case. Of these, 399 (19.1%) were plausible by our criterion of having a score greater than 4. The average number of plausible diagnoses per case was 11 with a standard deviation of 5, a maximum 22 and a minimum of 3. By chi-square test, the overall distribution of plausibility scores was statistically case-dependent (chi-square = 661.4, df = 210, p < .0001).

Research Question 2: Figure 1 displays the distribution of serial numbers of subjects generating the last of the plausible hypotheses, for 100 random orderings within each case and with results for all cases pooled. Each bar of this histogram corresponds to two serial numbers, so for example, the bar centered on a value of "21" includes serial numbers 21 and 22. With specific reference to Figure 1, the height of the bar centered on 21 is 150. This result can be interpreted as follows: In 150 subject orderings (of the 3600 orderings generated and analyzed), the final plausible diagnosis was generated by the 21st or 22nd subject (of the 36 subjects who worked each case).

The average serial number of the subject generating the final plausible diagnosis was 27.6, with a standard deviation across cases of 8.1. The curved line running through the figure represents the normal distribution with the mean and variance estimated from the data.

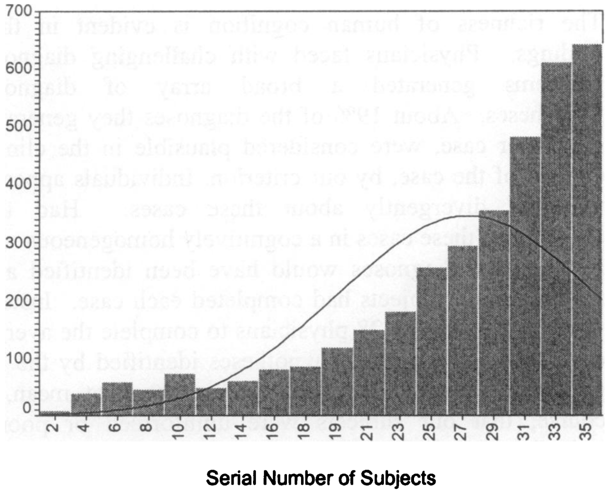

Figure 1. Distribution of the Serial Numbers of the Subjects Generating the Final Plausible Diagnosis, with 100 Orderings for Each of 36 Cases.

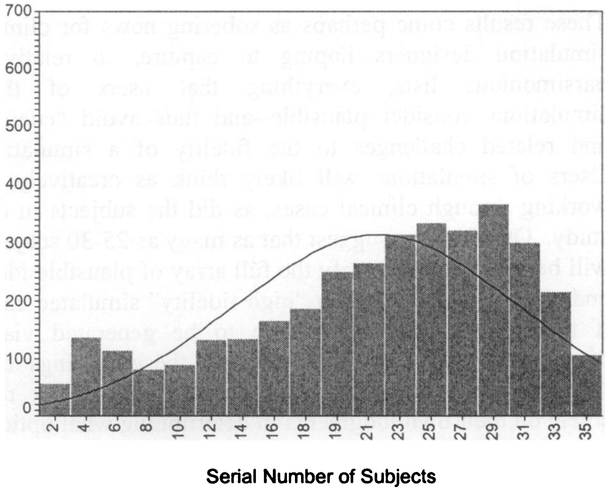

Figure 2. Distribution of the Serial Numbers of the Subjects Generating the Next-to-Last Plausible Diagnosis, with 100 Orderings for Each of 36 Cases.

Figure 2 (on the previous page) displays the analogous result exploring the serial number of the subject who generated the next-to-last plausible hypothesis. Again using the bar centered on 21 as an example, the height of the bar in Figure 2 is 270. So for 270 orderings (of 3600), the next-to-last plausible diagnosis was generated by the 21st or 22nd subject (of 36). The average serial number of the subject generating the next-to-last plausible diagnosis was 22.0, with a standard deviation across cases of 8.8. Although not shown in figures, the average serial number of the subject generating the "last-but-two" of the plausible diagnoses is 18.1 with a standard deviation across cases of 8.4.

DISCUSSION

The richness of human cognition is evident in these findings. Physicians faced with challenging diagnostic problems generated a broad array of diagnostic hypotheses. About 19% of the diagnoses they generated, or 11 per case, were considered plausible in the clinical context of the case, by our criterion. Individuals appeared to think divergently about these cases. Had they approached these cases in a cognitively homogeneous way, all plausible diagnoses would have been identified after relatively few subjects had completed each case. Instead, it was necessary for 28 physicians to complete the average case before all plausible hypotheses identified by the full group of 36 were generated. This does not mean, of course, that our subjects were uninformed or poorly-trained in their professions. It more likely illustrates that the educated mind is a fertile source of creative ideas. Faced with a challenging problem such as the diagnosis of a complex medical case, trained physicians asked to independently generate ideas will collectively generate a large number of them.

These results come perhaps as sobering news for clinical simulation designers hoping to capture, in relatively parsimonious lists, everything that users of their simulations consider plausible--and thus avoid "cueing" and related challenges to the fidelity of a simulation. Users of simulations will likely think as creatively, in working through clinical cases, as did the subjects in our study. Our findings suggest that as many as 25-30 subjects will be required to identify the full array of plausible ideas and options to include in a "high fidelity" simulated case, if these ideas and options are to be generated via a prospective study. It follows from these findings that simulation designers and case authors should not rely solely on their own judgments in determining what options should be included in a simulation.

One feature of our study that may modify this recommendation was the limitation of each subject to generating six hypotheses per case. If each subject were allowed to generate a longer list, it is possible that fewer subjects would be required to identify all plausible hypotheses, but this is by no means certain. For example, 80% of our subjects, when given the option of identifying up to six hypotheses, in fact identified fewer than six. The collective span of the group's thinking appears to greatly exceed the span of any individual.

Because the perfect is sometimes the enemy of the good, we examined the implications of being satisfied to identify all but one or two of the plausible hypotheses, in lieu of the full set. The numbers of subjects required to meet these relaxed criteria did diminish substantially. If simulation designers are willing to accept the risk of failing to identify some plausible options within a case, they could conduct prospective studies with fewer subjects than otherwise would be required. Along similar lines, we chose an intermediate threshold for considering a specific diagnosis to be plausible. Adopting a more inclusive threshold would, in general, increase both the number of plausible diagnoses and the number of subjects required to identify all or most of them. Adopting a more exclusive threshold would have the opposing effect.

It is evident from these findings that all phenomena we studied are highly case dependent. For example, the number of plausible hypotheses per case varied from 3 to 22 over the 36 cases included in our study. This carries a clear implication to simulation designers that not all disease domains can be approached in the same way. Identification of plausible options may be possible with relatively few subjects for some domains, and will require many more subjects for others.

Our decision to eliminate from the analysis the diagnostic hypotheses of medical students carries some implications. As stated earlier, the medical students diagnosed very few of these cases correctly and generated relatively few plausible hypotheses. Including them would have inflated estimates of the numbers of subjects required to generate the full set of plausible hypotheses. However, if simulations are designed for students, perhaps some implausible hypotheses generated consistently by students should be included in picklists and other components of the simulation. Exploration of this question goes beyond the scope of the present study, but it remains an issue for simulation design.

While the design of clinical simulations will always be part science and part art, the results of this study may be helpful in enhancing the scientific basis of this complex and challenging process.

REFERENCES

1. Issenberg SD, McGaghie WC, Hart, IR, Mayer JW, Felner JM et. al. Simulation technology for health care professional skills training and assessment. Journal of the American Medical Association 282: 861-866, 1999.

2. Dawson SL, Kaufman JA. The imperative for medical simulation. Proceedings of the IEEE 86: 479-483, 1998.

3. Friedman CP. The marvelous medical education machine or how medical education can be 'unstuck' in time. Medical Teacher 22: 496-502, 2000.

4. "Using Simulation for Education and Assessment". Conference sponsored by the University of Rochester Medical Center. May 4-6, 2002. See: www.urmc.rochester.edu/cpe/NEGEA.

5. Friedman CP. Anatomy of the clinical simulation. Academic Medicine 70: 205-209, 1995.

6. Harless WG, Drennon GC, Marxer JJ, Root JA, Miller GE. CASE: A computer-aided simulation of the clinical encounter. Journal of Medical Education 46: 443-448, 1970.

7. Friedman RB, Korst DR, Schultz JV, Beatty E, Entine S. Experience with the simulated patient-physician encounter. Journal of Medical Education 53: 825-836, 1978.

8. Friedman CP, Elstein AS, Wolf FM, Murphy GC, Franz TM, Heckerling PS, Fine PL, Miller TM, Abraham V. Enhancement of clinicians' diagnostic reasoning by computer-based consultation: A multisite study of 2 systems. Journal of the American Medical Association 282: 1851-1856, 1999.

9. Friedman, C.P., Elstein, A.S., Wolf, F., Murphy, G., et. al. Measuring the quality of diagnostic hypothesis sets for studies of decision support. Ninth World Congress on Medical Informatics, 864-868, 1998.

ACKNOWLEDGEMENTS

This work was supported by grant R01-LM-05630 from the National Library of Medicine, and also by Assessment Technologies, Inc., a subsidiary of the American Board of Family Practice. The authors wish to acknowledge the contributions of Drs. Arthur Elstein and Fredric Wolf for their efforts in collecting and analyzing the original dataset on which this study is based. They also wish to acknowledge the advice of Drs. Michael Hagen and Walton Sumner of Assessment Technologies, as well as that of Dr. Stephen Downs of Indiana University.